(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（Ａ）**　(11)特許出願公開番号

特開平6－274193

(43)公開日　平成６年(1994)９月30日

(51)Int.Cl.[5]	識別記号	庁内整理番号	ＦＩ	技術表示箇所
Ｇ１０Ｌ　3/00	５３１　Ｇ	9379－5H		
	５１５　Ｂ	9379－5H		
Ｇ０６Ｆ　15/40	５３０　Ｖ	9194－5L		

審査請求　未請求　請求項の数４　ＦＤ　（全　８　頁）

(21)出願番号　特願平5－84154

(22)出願日　平成５年(1993)３月17日

(71)出願人　000006655
新日本製鐵株式会社
東京都千代田区大手町２丁目６番３号

(72)発明者　柴田　克信
相模原市淵野辺５－10－１　新日本製鐵株式会社エレクトロニクス研究所内

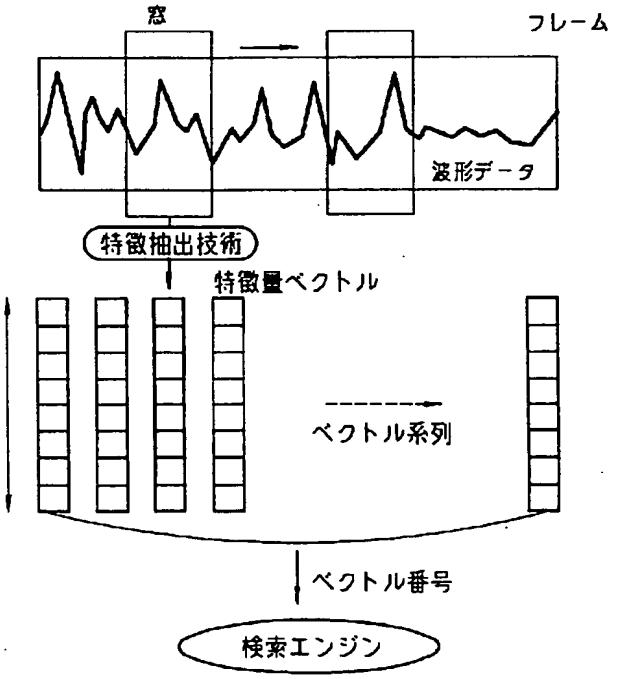
(74)代理人　弁理士　國分　孝悦

(54)【発明の名称】　データベース検索システム

(57)【要約】

【目的】　波形または数値列のデータについて特徴量抽出、量子化を行うデータベース検索システムを提供する。

【構成】　検索対象となる波形・数値列データを複数のブロックたるフレームに分割し、通し番号（フレーム番号）を付与する。このフレームを検索対象データの検索単位とし、学習においては検索単位ごとに特徴量抽出、量子化、コード列化の処理を施し、データを保存する。特徴量抽出は窓付きのフーリエ変換、ウェーブレット変換、一般の直交変換等により行われる。検索においては、検索キーとなる波形・数値列データに対して、特徴抽出、量子化、コード列化、重み付けの処理が行われる。



【特許請求の範囲】
【請求項１】　波形、数値列データを対象とするデータベース検索システムにおいて、
波形・数値列データを検索を行う所定の検索単位に分割する分割手段と、
前記分割手段によって分割された前記検索単位ごとに特徴量抽出を行う特徴量抽出手段と、
前記特徴量抽出手段によって特徴量抽出を行われたデータについて量子化を行う量子化手段と、
検索時に、検索キーとなる波形・数値列データに対して、重み付けの処理を行う重み付け手段とを具備することを特徴とするデータベース検索システム。
【請求項２】　前記特徴量抽出手段は、データの時系列を第一の軸とし、各特徴成分を第二の軸とする二次元平面上に特徴量の分布として、特徴量の抽出を行うことを特徴とする請求項１に記載のデータベース検索システム。
【請求項３】　検索対象の物件毎にその近傍特徴量を記憶した記憶手段と、
検索キーの近傍特徴量と検索対象の上記近傍特徴量との合致度を物件毎に求め、物件番号を合致度の降順に出力する検索手段とを具備するデータベース検索に用いられることを特徴とする請求項１のデータベース検索システム。
【請求項４】　検索対象のｉ番目の物件のｊ番目のデータ列$C_{i,j}$に関する量子化量ｘとその近傍のｋ個のデータ列$C_{i,j+1}, C_{i,j+2}, \ldots, C_{i,j+k}$に関する量子化量
ｙとを
$x = f(C_{i,j})$
$y = g(C_{i,j}, C_{i,j+1}, C_{i,j+2}, \ldots, C_{i,j+k})$
によって求め、得られたｘ、ｙの値に基づいて定められる記憶手段の位置にその物件の通番ｉを記憶するデータベース検索に用いられることを特徴とする請求項３のデータベース検索システム。
【発明の詳細な説明】
【０００１】
【産業上の利用分野】本発明は、データベースから必要な情報を取り出すためのデータベース検索システムに関し、特に波形、数値列等のデータについてのデータベース検索システムに関する。
【０００２】
【従来の技術】現状のデータベース検索における検索空間圧縮の手法としては、キーワードなどのインデックス情報を付加する方式によるものが一般的である。また、対象物件数が比較的小数のケースでは、全物件検索方式が実用化されている。たとえば文書データにおいては効率的な全物件検索手法として、ボイヤーモア法が考案されている。
【０００３】
【発明が解決しようとする課題】このインデックス検索方式は、以下のような欠点を有する。
（１）物件１つ１つにインデックスを付加しなくてはならない。
（２）任意のインデックスを付加して行くと、その個数は膨大になるため、たとえばキーワードインデックスの場合にはシソーラスによる管理などを必要とし、その維持に多大なコストがかかる。
（３）付加されるインデックスが必ずしも適切なものとは限らない。すなわち現状のデータベース検索方式では、特に物件数が膨大になった時に必要なコストに比してパフォーマンスが伸びない傾向が現れる。
【０００４】一方、全物件検索方式では、上記のような問題は発生しない。しかし、直接検索方式では、物件数が膨大になったとき、検索時間は対話的な時間の域を大幅に超過し、実用にならないのが現状である。また、全物件検索方式におけるあいまい一致条件では完全一致よりさらに検索時間が必要となる。
【０００５】出願人は先に、全物件検索でありながら、検索時間を飛躍的に短縮することができ、またあいまい一致検索ができるデータベース検索方式を特願平３－１２２７６６号として提案した。
【０００６】ところで、上述のボイヤーモア法による全物件検索では、文書以外のデータ、たとえば物理的な時系列データを扱うことができないという問題があった。また、前記の特願平３－１２２７６６号に係る検索システムにおいても、波形、数値列等のデータの場合には、これらのデータ列の有意な情報単位が文書データと異なり、特徴量を抽出し量子化することが困難であるため、これらのデータを対象とすることができないという問題があった。
【０００７】すなわち、波形、数値列データは、文書データ等と異なり、サンプリングされた数個の値だけで特徴をもつことがない。したがって、何らかの方法によって情報の抽出度を上げて検索データへの依存度を弱める必要がある。現状において、波形、数値列データの検索は主にDynamic Programingマッチング、隠れマルコフモデルなどの手法を用いた逐次マッチングにより実現されている。しかしながら、これらの手法は多くの計算コストを必要とするため、特に大規模システムへの適用においては検索時間の点で問題を生じる。また、文書データにおけるキーワードに相当する２次情報を事前に作成することで検索時間を短縮する手法も考えられるが、有意な情報単位の一般的定義が困難であるため、実現された例は少ない。
【０００８】本発明は、上記のような問題に鑑み、特定の意味を有する情報単位が文書データと比較して大きい波形、数値列データについて、全物件検索でありながら検索時間を飛躍的に短縮することができ、またあいまい一致検索ができるデータベース検索方式を提供することを目的とする。

【０００９】
【課題を解決するための手段】本発明の波形、数値列データを対象とするデータベース検索システムは、波形・数値列データを検索を行う所定の検索単位に分割する分割手段と、分割手段によって分割された検索単位ごとに特徴量抽出を行う特徴量抽出手段と、特徴量抽出手段によって特徴量抽出を行われたデータについて量子化を行う量子化手段と、検索時に、検索キーとなる波形・数値列データに対して、重み付けの処理を行う重み付け手段とを具備する。
【００１０】
【作用】本発明によれば、波形、数値列データ情報の抽象度を上げて検索データへの依存性を弱めることができ、波形、数値列データについて全物件検索でありながら検索時間を飛躍的に短縮することができ、またあいまい一致検索ができる。
【００１１】
【実施例】次に図面により本発明の実施例を説明する。本発明の検索システムにおいては、検索対象となる波形・数値列データを検索を行う単位である検索単位に分け、この検索単位ごとに昇順の符合を付与する。学習においては１つの検索単位に対して特徴抽出、量子化、コード列化の処理を施し、データを保存する。
【００１２】検索においては、検索キーとなる波形・数値列データに対して、特徴抽出、量子化、コード列化、重み付けの処理が行われる。
【００１３】本発明による検索システムにおいて、対象とされる波形、数値列データの特徴抽出は、データが時系列であると考えた場合、時間的に局所化した関数との内積を取ることにより特徴量を抽出することが基本的な考え方である。内積値は、その時間近傍での特徴を反映していると考えることができる。特徴量の抽出は、以下の方法が例として挙げられる。
１．窓付きのフーリエ変換、特徴軸は周波数
これは窓関数を用いて境界付近の歪みを抑えつつ、一定の分解能で短時間の周波数成分を抽出するものである。
２．ウェーブレット変換、特徴軸は基本関数のスケール
これは時間と周波数の成分を同時に扱い、周波数によって時間分解能が変化するものである。
３．一般の直交関数系、特徴軸は基本多項式の展開係数
これは時間分解能一定で、波形のトレンドに重点を置くものであり、たとえばルジャンドル多項式があげられる。
上記の特徴抽出は十分に高速であることが、望ましい。
【００１４】次にこれらの特徴抽出方法について説明する。図１に示すように、波形、数値列データを複数のブロックたるフレームに分割し、通し番号（フレーム番号）を付与する。このフレームを検索対象データの単位とする。
【００１５】次に波形、数値列データを時間軸方向、すなわち波形の振幅方向と垂直な軸方向に窓を被せ、この窓を特徴量抽出対象領域としてこの区間で特徴抽出を行う。
【００１６】この窓を走査し、各区間で計算された特徴ベクトル、すなわち特徴抽出の方法にしたがって周波数、スケール、展開係数を成分とする特徴ベクトルの時系列を生成する。
【００１７】例を上げて説明する。サンプリングされた波形あるいは数値列データを時間ｔの関数であるとし、Ｆ(t)（t=0,1,,,,）で表す。時間０の回りに局在した関数をＧ(t,p) とする。ただし、ｐは特徴を定義するパラメータである。このとき時刻Ｔの近傍の特徴値は内積
【００１８】
【数１】

$$I(T, p) = \sum_{t=T-a}^{T+a} F(t) \times G(t-T, P)$$

【００１９】で定義される。ただし、２×ａはＴを中心とする窓の区間の大きさを表す。すなわち窓は区間[T-a, T+a]で定義される。
【００２０】例えば、Ｇ(t, p)は窓付きのフーリエ変換の場合、
G(t, p) = 0.5× exp(-i ×p ×t) {1+ cos(π×t/a)}　|t |≦ a
(i は虚数単位)　$i^2 = -1$
G(t, p) = 0　|t |≧ a
ウェーブレット変換の場合、
G(t, p) = exp(-t ^2/p ^2+i ×5 ×t/p)/ √p
　　|t |≦ a
(i は虚数単位)　$i^2 = -1$
G(t, p) = 0　|t |≧ a
などを用いることができる。
【００２１】また、関数Ｇ()が複素関数である場合にはI(T,p)の絶対値を特徴量とし、T-pの二次元平面上に特徴量が計算される。
【００２２】図２（ａ）（ｂ）には、窓付きのフーリエ変換およびウェーブレット変換の基本関数の例をそれぞれ示す。また、図３（ａ）（ｂ）には、窓付きのフーリエ変換およびウェーブレット変換の時間分解能の比較を示す。
【００２３】次に図１に示すように、得られた特徴ベクトルの量子化を行う。量子化はまず、軸方向Ｔ 、ｐ の量子化が行われる。
【００２４】まず、各軸方向で区間を設定する。例えば、Ｔ 軸方向で[a×k,a ×(k+1)](k = 0,1,,,,) 、ｐ 軸方向で[b×l,b ×(l+1)](l = 0,1,,,,) とする。この区間内で量子化値 D[i,j] を所定の演算により決定する。例えば

1.　区間ごとにある規則に従って代表点を選び、その点での特徴量をそのままその区間の特徴量とする。
2.　区間内での平均値を計算する。
などの方法が考えられる。これらにより特徴平面はベクトル系列（あるいは行列）に変換される。

【００２５】次に、特徴量の量子化として、それぞれのベクトル成分を最大値などで正規化した後に量子化する。例えば４ビット、２ビットなどで表現される最大数で正規化し、このビットで量子化する。

【００２６】上記の処理でデータの正規化は一応なされたが、さらにこのベクトル近傍（ベクトル同士の近傍あるいは、成分での近傍）から数値列を再定義し近傍特徴量とすることも可能である。検索単位である波形のｋ番目のベクトルのｌ番目の要素をV[k,l]とする時、例えば、この近傍での特徴量 $I_R[k,l]$ は、関数h()を定義して、

$I_R[k,l] = h(V[k,l], V[k,l+1], V[k+1,l], V[k+1,l+1])$

で求められる。

【００２７】一方、検索時には、近傍特徴量に特徴軸方向に重み付けを行うことも可能である。これは、検索時に任意に設定する、あるいは類似であると定義するものを繰り返し提示し、学習することによって定義することも可能である。

【００２８】例えば、特徴量 $I_R[*,l]$（* は任意）での重みをA[ｌ]とし、波形１の特徴量を $I_{R1}[k,l]$、波形２の特徴量を $I_{R2}[k,l]$ とする時、$I_{R1}[k_1,l] = I_{R2}[k_2,l]$ となるような k_1, k_2 が存在するならば、A[ｌ]+=dAとし、それ以外のｌについては、A[ｌ]-=dAとする。ただし、A[l]の初期値は１とし、dAは１に比べ非常に小さい数である。

【００２９】上記のような波形、数値列データの特徴量抽出、量子化は、たとえば次のようなデータ検索システムにおけるデータの処理に適用できる。

【００３０】図４は、本発明が適用される自己相関記憶型パターン検索システムのデータフロー図である。この検索システムでは、予め全対象物件から事象（情報）の位相情報を全て捨象した近傍特徴量データを作成し、そのデータ群に対して全物件検索を行なう。検索のアルゴリズムは、学習ステップと検索ステップとからなる。学習ステップでは、物件毎に近傍特徴量行列が位相情報として作成される。検索ステップでは、検索キーと近傍特徴量行列とのマッチング演算が行なわれ、物件ごとにマッチング度（類似度）を示す評価結果を得る。以下、各ステップについて説明する。

【００３１】（１）、学習ステップ
図４に於いて、検索対象１０は、例えば日本語、英語、ドイツ語、フランス語、ヘブライ語、ロシア語などの文書データ、或いは本発明の特徴たる量子化された波形数値データや、化学構造式、遺伝子情報などである。このような検索対象に対して、まず正規化手段Ｓ１により正規化の処理を行なう。一般に検索対象は、情報の最小単位（文書であればアルファベットなどの文字、数値チャートであれば、ある時刻における実数値など）の列で表現されている。それをなんらかの方法でｎ階調の整数列に変換する。これをデータの正規化と呼ぶ。本発明においては前述のように正規化が行われる。

【００３２】正規化されたデータ２０は、次に学習手段Ｓ２により近傍特徴量行列３０の形式に畳込まれる。ここで近傍特徴量をとる演算式は種々考えられる。この演算式は検索の鋭さ（過検出の少なさ）にも影響を与える。

【００３３】今、ｉ番目の物件のｊ番目のデータを $C_{i,j}$ とし、$C_{i,j}$ に関する量子化量ｘと $C_{i,j}$ の前方ｋ近傍に関する量子化量ｙを次のようにして求める。ここでは、検索される対象物件がｎ個あるとし、そのうちのｉ番目の物件の量子化について説明する。ｉ番目の物件において、図５に示すように正規化された数値列135,64,37,71,101,...が並んでいるとすると、$C_{i,j}$ に関する量子化量ｘは、

$x = f(C_{i,j})$

$C_{i,j}$ の前方ｋ近傍に関する量子化量ｙは

$y = g(C_{i,j}, C_{i,j+1}, C_{i,j+2}, \ldots, C_{i,j+k})$

で求められる。

【００３４】ここで、$f(C_{i,j})$ は $C_{i,j}$ に関するｎ段階量子化関数である。すなわち、ｉ番目の物件のｊ番目のデータ $C_{i,j}$ について所定の演算を行って得られる値であり、１～ｎのいずれかの整数で表される。したがって、得られたｘの値によって図６に示す行列（座標）においてｘ軸方向の位置が１～ｎの範囲で定まる。

【００３５】また、$g(C_{i,j}, C_{i,j+1}, C_{i,j+2}, \ldots, C_{i,j+k})$ は、$C_{i,j}$ の前方ｋ近傍に関するｍ段階量子化関数である。すなわち、ｉ番目の物件のｊ番目のデータ $C_{i,j}$ とそのデータの近傍の所定の数のデータについて所定の演算を行って得られる値であり、１～ｍのいずれかの整数で表される。たとえば図５に示すようにｊ番目のデータ $C_{i,j}$ が１３５であり、ｋが３の場合には、$C_{i,j+1}, C_{i,j+2}, C_{i,j+3}$ としてデータ１３５に続くデータ６４、３７、７１を抽出し、これらのデータとデータ１３５との相関について所定の演算を行う。ｊ番目のデータ $C_{i,j}$ が次の６４の場合には、$C_{i,j+1}, C_{i,j+2}, C_{i,j+3}$ としてデータ６４に続くデータ３７、７１、１０１を抽出し、これらのデータとデータ６４との相関について所定の演算を行う。

【００３６】このようにして得られたｙの値によって、図６に示す行列（座標）におけるｙ軸方向の位置が１～ｍの範囲で定まる。したがって、上記のようにｘ、ｙを求めることによって図６に示す行列（座標）における位置が定まる。

【００３７】本システムでは、各物件情報は、上記のようにして求めたｘ、ｙに対して物件の通番ｉと重みｗ

(x,y,i)の組として記憶される。重みｗ（x,y,i)は、データｘ、ｙ、ｉから所定の演算によって求められるが、通常は重みｗ（x,y,i)の値は１に固定される。

【００３８】上記のようにして求められたデータ$C_{i,j}$ごとにｘ、ｙの値に基づき図６に棒によって示されるように、データを記憶する。すなわち、データ$C_{i,j}$のｘ、ｙの値によって定められる座標の位置に、その物件の通番ｉとその重みｗ（x,y,i)を組みとしたデータを記憶する。同図ではこのようなデータが記憶されるごとに棒の長さが延びるように表されている。通常は重みｗ（x,y,i)は１とされるから、物件の通番ｉのデータのみがｘ、ｙの値によって定められる座標の位置に記憶されてゆく。

【００３９】この様にして作成された近傍特徴量行列に物件の識別番号を付加して構造ファイル４０として保存する。

【００４０】（２）、検索ステップ

まず検索キー５０を入力する。この検索キー５０に対して学習ステップと同一の正規化方法に基づく正規化手段Ｓ３によりキー情報を整数列に正規化する。

【００４１】次に、検索手段Ｓ４において、学習ステップと同一の自己相関計算式f()、g()を用いて各物件に対応する正規化された数値列の先頭からｘ、ｙの組の系列を作成する。次に、このｘ、ｙの組の系列に基づいて、物件ｋに対する検索キーの含有度数ω_kとして、Ｖ（x_j, y_j, k）をｊ＝１～ｍについて合計することにより算出する。

【００４２】ただし、Ｖ（x_j, y_j, k）は、物件情報リストが物件ｉについての重みを持つ場合、はその重みに等しく、持たない場合には０と定める。

【００４３】したがって、検索すべき数値列のｘ、ｙの組に対応する図６のｘ、ｙの位置にデータがある場合（棒がある場合）には、別に設けられた記憶手段のそのデータに示される物件の通番ｉの格納箇所にその重みの値を記憶させる。

【００４４】次に、評価結果出力手段Ｓ５において、物件毎に得られた構造評価値score（合致度）を完全一致の場合の評価値で割って、検索キーの含有確率を求め、評価結果のリスト７０を得る。更にソート手段Ｓ６において、このリスト７０を含有確率の降順にソートしソート済みリスト８０を得る。

【００４５】このソート済みリスト８０が検索結果であり、その上位物件を参照することにより、検索キーが物件中に含まれている確率が高い物件名を知ることができる。含有確率は、完全一致及び不完全一致の全てについて求まるから、あいまい一致検索を行なうことができる。

【００４６】また、検索キーの全情報についての全物件探索であるから、検索もれが発生する確率は、本質的に零であると言う特徴がある。

【００４７】また、１つの物件に対する検索キーの評価時間は、キーのデータ数のみに依存し、物件の大きさには依存しない。従って、非常に高速に検索を行なうことができる。

【００４８】また検索結果のリストどうしの論理演算を行うことにより、検索条件に対するＡＮＤ、ＯＲなどの検索演算処理も高速に実行できる。式（１）の自己相関式は上述の例の他に種々考えることができる。例えば、

f: x→x

g: (x,y)→x-y （または｜x-y｜）

とすれば、隣接データ及び一つ置きのデータの差分（または差分の絶対値）を相関情報として近傍特徴量行列を作ることができる。また幾つかのデータ列の個々のデータ整数値に対し四則演算を施すことにより近傍特徴量を取り出してもよい。

【００４９】近傍特徴量は、各物件の全データを対象とし取り出さなくてもよい。例えば、物件データ中の特定の一つまたは一つ以上の整数値、特定の範囲の整数値、或いはデータ列を構成する各バイト中の特定の１つまたは一つ以上のビットを除外して近傍特徴量を捨象してもよい。

【００５０】上述の例では、近傍特徴量によって生成される行列は、２５６次のビット行列であり、これは８Ｋバイトに相当する。従って、１物件のデータが１Ｋバイト程度であるデータベースでは、効率のよいシステムであるとは言えない。そこで上記のようなデータ圧縮手段Ｓ７を設けてデータ圧縮を行なって構造ファイル４０の容量を減らすのがよい。

【００５１】上述の実施例において，正規化手段Ｓ１、学習手段Ｓ２、正規化手段Ｓ３、検索手段Ｓ４、評価結果出力手段Ｓ５、ソート手段Ｓ６、データ圧縮手段Ｓ７は、コンピュータプログラムによって構成することができるが、論理回路素子を用いて専用のハードウエアを構成してもよい。

【００５２】前述のような波形、数値列データの特徴量抽出、量子化を上記の検索システムに適用すれば、波形、数値列データの検索を有効に行うことができる。

【００５３】

【発明の効果】本発明は波形、数値列データの特徴量抽出、量子化を行うようにしているからこれらのデータを対象とする検索を有効に行うことができる。

【図面の簡単な説明】

【図１】本発明による特徴量抽出、量子化の例を示す図である。

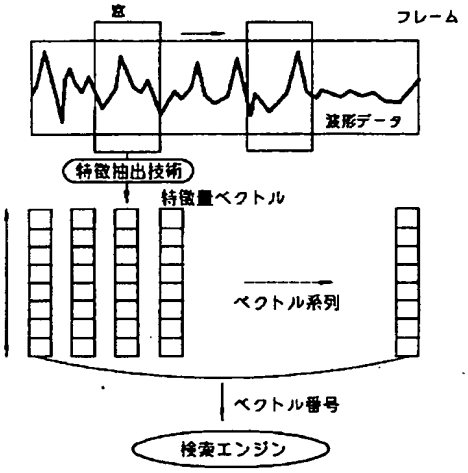
【図２】本発明に用いられる基本関数の例を示す図である。

【図３】本発明に用いられる変換の解像度の例を示す図である。

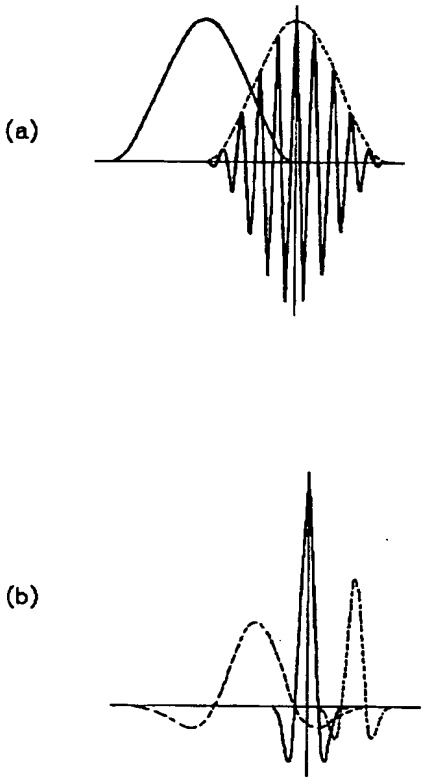
【図４】本発明によるデータベース検索システムのデータフロー図である。

【図５】近傍情報の量子化を示す図である。
【図６】記憶される情報構造を示す図である。
【符号の説明】
３０　近傍特徴量行列
４０　構造ファイル
５０　検索キー
６０　正規化キー
７０　評価結果リスト
８０　ソート済みリスト
Ｓ１　正規化手段
Ｓ２　学習手段
Ｓ３　正規化手段
Ｓ４　検索手段
Ｓ５　評価結果出力手段
Ｓ６　ソート手段
Ｓ７　データ圧縮手段

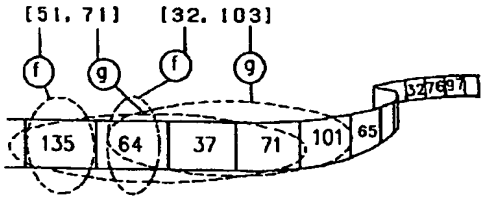
【図１】



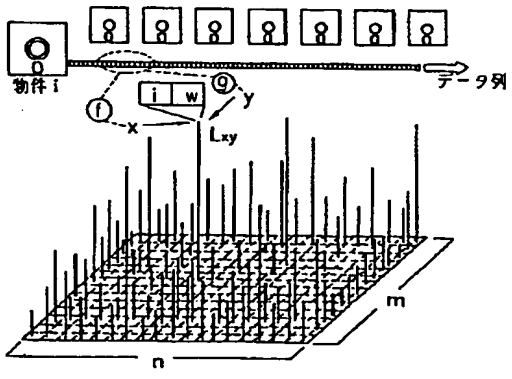
【図２】



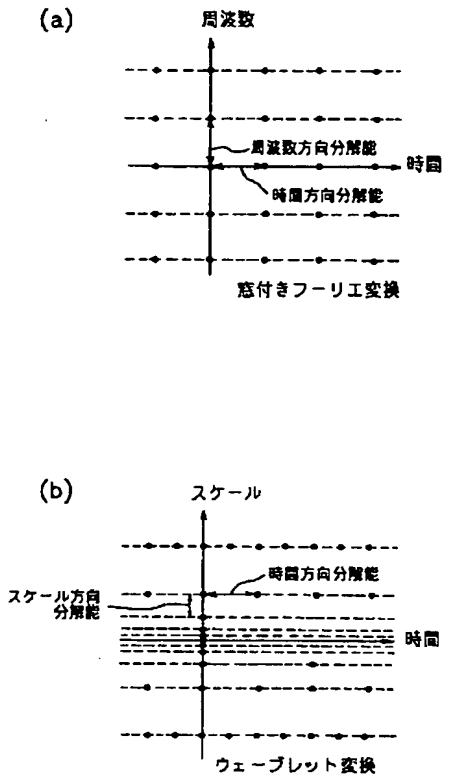
【図５】



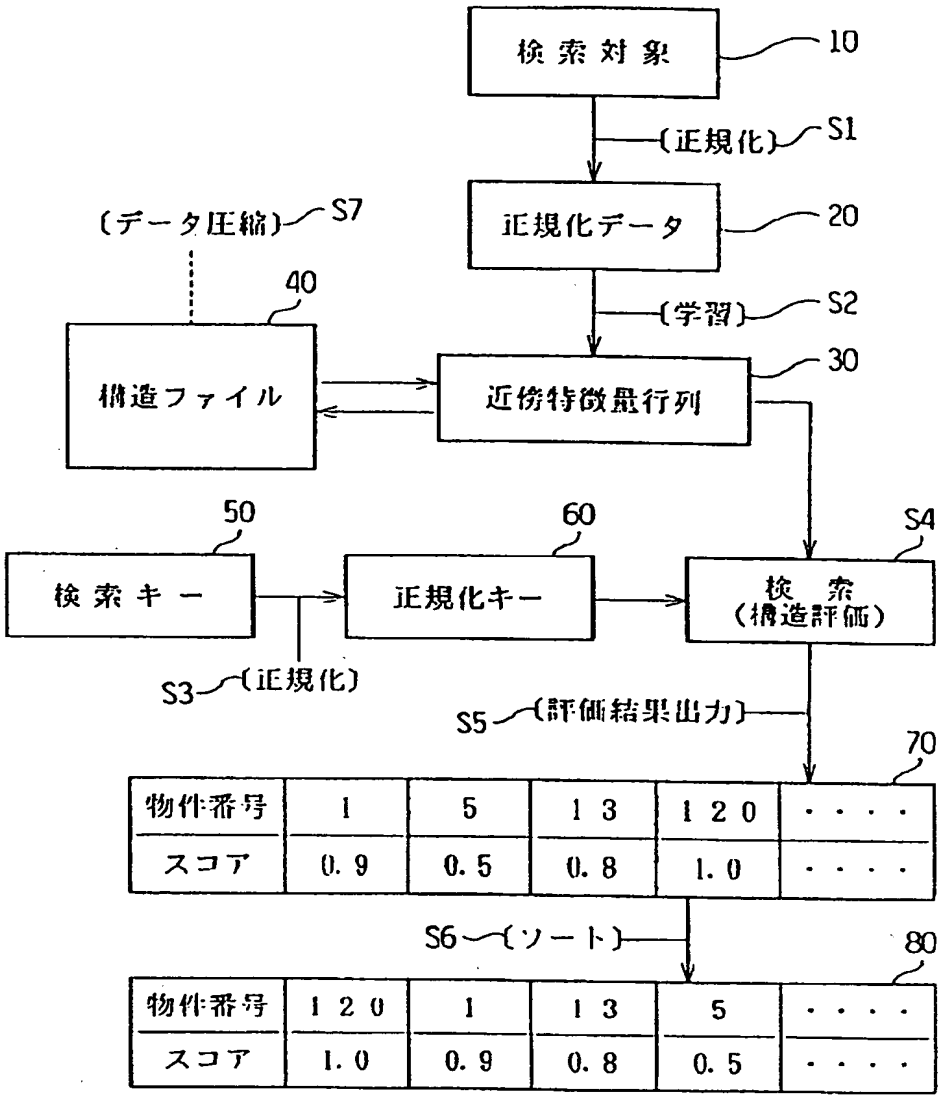
【図６】



【図３】


(a)
周波数
周波数方向分解能
時間
時間方向分解能
窓付きフーリエ変換
(b)
スケール
時間方向分解能
スケール方向分解能
時間
ウェーブレット変換

【図４】


検索対象
10
(正規化)
S1
正規化データ
20
(データ圧縮)
S7
40
構造ファイル
(学習)
S2
近傍特徴量行列
30
50
検索キー
60
正規化キー
S4
検索
(構造評価)
S3
(正規化)
S5
(評価結果出力)
70
物件番号 | 1 | 5 | 13 | 120 | ・・・・
スコア | 0.9 | 0.5 | 0.8 | 1.0 | ・・・・
S6
(ソート)
80
物件番号 | 120 | 1 | 13 | 5 | ・・・・
スコア | 1.0 | 0.9 | 0.8 | 0.5 | ・・・・